International Journal of Applications of Fuzzy Sets and Artificial Intelligence (ISSN 2241-1240), Vol. 8 (2018), 143 - 154

Decision Trees in Data Mining

Abdel-Badeeh M. Salem [1], Michael Gr. Voskoglou [2]

[1] Faculty of Computer & Information Sciences
Ain Shams University, Cairo, Egypt
absalem@cis.asu.edu.eg , abmsalem@yahoo.com

[2] School of Technological Applications
Graduate Technological Educational Institute of Western Greece
voskoglou@teiwest.gr , mvosk@hol.gr

Abstract

Information mining is the process of acquiring knowledge from patterns discovered by extracting (mining) from information granules. Data mining, a special case of information mining, is the computing process of collecting and summarizing data from a web site's hyperlink structure, page content, or usage log in order to identify patterns in large data sets. The paper at hands reviews the decision tree algorithms for data mining, the most popular being the C4.5 and the CART. Our conclusions are also stated on comparing the advantages and disadvantages of the above two data mining algorithms and their applications are discussed.

Keywords: *Decision Trees (DT), Information and Data Mining (IM and DM), Machine Learning, DT Learning Algorithms, C4.5,* C*lassification* and R*egression* T*ree* (*CART*).

Received: May 12, 2018

1. Introduction

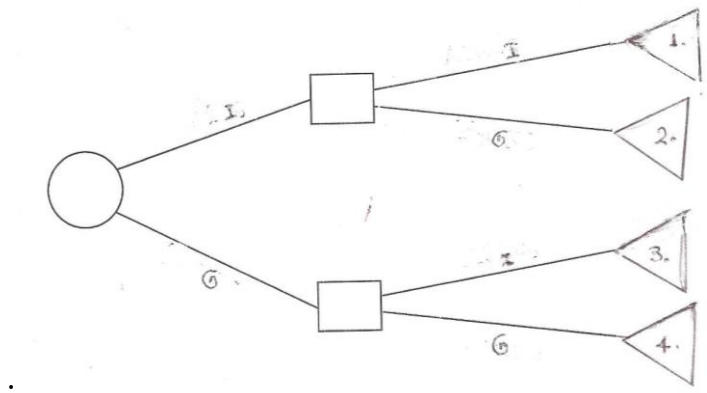
Decision Trees (DTs) are commonly used in Operations Research, specifically in ***Decision Analysis***, to help identify a strategy for reaching a goal. It is recalled that a DT is a tree – like graph representing the logical structure of a decision problem and playing the role of a visual decision support tool for calculating the expected

values of the corresponding alternatives. In particular, DTs are very useful in complicated problems in which successive decisions are made step by step.

Drawn from left to the right a DT has splitting ***branches*** starting or ending with three types of ***nodes***:

- ***Decision nodes***, denoted by squares.
- ***Chance nodes***, denoted by circles, and
- ***End nodes,*** denoted by triangles

A typical example of a DT is that of Figure 1, representing a judge's decision about an accused person.



:

Figure 1: DT representing a judge's decision

In the above DT "I" stands for "Innocent" and "G" stands for "Guilty", whereas the final outcomes given by the numbers written inside the corresponding end nodes are: 1 = An innocent person has been decided to be innocent, 2 = An innocent person has been decided to be guilty, 3 = A guilty person has been decided to be innocent and 4 = A guilty person has been decided to be guilty.

Among the other decision support tools, DTs have several advantages:

- They are simple to understand and interpret.
- They help to determine worst, best and expected values for different scenarios and they allow the addition of new possible scenarios.

- They use a white box model whose internals can be viewed but usually cannot be altered. This helps in various ways; e.g. if a programmer can examine the source code, the weaknesses in an algorithm can be detected and fixed much easier.
- They can be combined with other decision techniques, etc.

On the other hand, one of the main disadvantages of the DTs is that they are unstable, in the sense that small changes to the given data could lead to big changes to the corresponding DT. Also the calculation of the outcomes in a DT is sometimes proved to be very complicated, especially when uncertain variables are involved.

Apart from the decision analysis, DTs have been also widely used in different domains as a robust ***Machine Learning*** tool **[1]**. It is recalled that machine learning is the branch of Computer Science that uses statistical techniques to give computers the ability to "learn", i.e. to progressively improve performance on a specific task, with data not being explicitly programmed **[2, 3]**. ***DT Learning,*** which uses a DT to go from observations about an item to conclusions about the item's target value, is one of the predicting modeling approaches utilized in Statistics, and ***Data Mining (DM)***.

In this work the use of DTs in DM paradigms is studied. The rest of the paper is organized as follows: In Sections 2 the definition and the main modes of ***Information Mining (IM)*** are presented including DM as a special case. Section 3 is devoted to the description of the main DM algorithms, among which the ***C4.5 Tree*** and the **C*lassification* and *Regression* T*ree*** (***CART***) are the most popular. The applications of those two algorithms are discussed in Section 4 Finally, in Section 5 our conclusions are drawn and some hints are given for future research on the subject.

2. Information Mining

IM is the process of acquiring knowledge from patterns discovered by extracting (mining) from data or information granules **[4]**.

Figure 2: Modes of Information Mining

The most important modes of IM, which are represented graphically in Figure 2 include:

- ***Data Mining (DM),*** which is the computing process of collecting and summarizing data from a web site's hyperlink structure, page content, or usage log in order to identify patterns in large data sets. DM involves methods at the intersection of machine learning, statistics, and database systems **[5]**.
- ***Ethical DM***, which involves ethical issues in Web DM.
- ***Image Mining,*** which is the process of searching and discovering valuable information and knowledge in large volumes of data. Image mining draws basic principles from concepts in databases, machine learning, statistics,

pattern recognition and soft computing. Using DM techniques enables a more efficient use of data banks of earth observation data.

- ***Mobile Mining***, which is generally the idea of mining on a mobile phone. With conventional crypto currencies real and valuable mining is possible. Everyone is on mobile now.
- ***Structural representation mining,*** which is the process of finding and extracting useful information from semi-structured data sets. Graph mining, sequential pattern mining and molecule mining are special cases of structured data mining.
- ***Graph Mining,*** which studies the problem of discovering typical patterns of graph data, where the structure of the data is just as important as their content.
- ***Frequent Pattern Mining,*** i.e. the process of finding frequent patterns which plays an essential role in mining associations, correlations, and in many other interesting relationships among data. If a substructure occurs frequently in a graph database, it is called a frequent structural pattern.
- ***Social Network Mining,*** which is the process of representing, analyzing, and extracting actionable patterns and trends from raw social media data.
- ***Ontology Mining,*** which is directly motivated by the need for formalization of the data mining domain.
- ***Behavior Mining,*** which contributes to the in-depth understanding, discovery, applications and management of behavior intelligence.
- ***Audio Mining,*** a technique by which the content of an audio signal can be automatically analyzed and searched. It is most commonly used in the field of automatic speech recognition, where the analysis tries to identify any speech within the audio.
- ***Text Documents Mining***, which is roughly equivalent to text analytics, a process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning.

- ***Web Mining,*** which is the integration of information gathered by traditional DM methodologies and techniques with information gathered over the World Wide Web.
- ***Video Mining,*** which is the discovery of patterns in audio-visual content.

Next we shall focus on the DT algorithms developed for DM.

3. Decision Tree Algorithms for Data Mining

While in decision analysis a DT is used to visually represent decisions and decision making, in DM a DT describes data that can be an input for decision making. In DM, DTs can be described as the combination of mathematical and computational techniques to aid the description, generalization and categorization of a given set of data.

Data come in records of the form $(X, Y) = (x_1, x_2, \ldots\ldots, x_n, Y)$ **(1)**.

The dependent variable Y is the ***target variable*** that we are trying to understand, classify or generalize. The vector X is composed of the features (samples), x_1, x_2, ….., x_k, that are used for the corresponding task.

The DTs used in DM are of two main types:

- ***Classification trees***, where the target variable can take a discrete set of values.
- ***Regression trees***, where the target variable can take continuous values, typically in the form of real numbers (e.g. the price of a house, a patient's length of stay in a hospital, etc.).

The process of computing classification and regression trees can be characterized as involving four basic steps:

1. Specifying the criteria for predictive accuracy
2. Selecting splits
3. Determining when to stop splitting
4. Selecting the "right-sized" tree.

In general, trees used for regression and trees used for classification have some similarities, but also some differences, such as the procedure used to determine where to split. **[6]**.

Traditionally DTs are designed manually. However, often in complicated problems DTs can grow very big being hard to be drawn by hand and special software has been developed for use in such cases. There are many specific decision-tree algorithms for DM. Notable ones include:

1. ***ID3*** (Iterative Dichotomiser 3)
2. ***C4.5*** (successor of ID3)
3. ***CART*** (**C**lassification **A**nd **R**egression **T**ree)
4. ***CHAID*** (**CH**i-squared **A**utomatic **I**nteraction **D**etector), which performs multi-level splits when computing classification trees.
5. ***MARS***, which extends DTs to handle numerical data better.
6. ***Conditional Inference Trees***, a statistics-based approach that uses non-parametric tests as splitting criteria, corrected for multiple testing to avoid over-fitting. This approach results in unbiased predictor selection and does not require pruning.

According to a survey that took place in the IEEE International Conference on Data Mining (ICDM) in 2006, two DT algorithms were elected as the most popular data mining algorithms, C4.5 and CART, ranked with number one and number ten respectively among the other algorithms **[7]**.

C4.5 is an algorithm developed by ***Ross Quinlan*** in 1986 **[8 - 10]**, being actually an extension of Quinlan's earlier ID3 algorithm. C4.5 builds DTs from a set of training data in the same way as ID3, using the concept of ***information entropy H*** **[11]**. For this, if p_i is the probability of appearance of the feature x_i, i = 1, 2, …, n,

then $H = -\sum_{i=1}^{n} p_i \log_2 p_i$ **(2)**

At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of features into subsets enriched in one class or the other. The splitting criterion is the ***normalized information gain*** (difference in entropy)

[12]. The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurs on the smaller sub-lists.

Authors of the Weka machine learning software described the C4.5 algorithm as "a landmark DT program that is probably the machine learning workhorse most widely used in practice to date" **[13]**.

The term Classification And Regression Tree (CART) analysis is an umbrella term used to refer to both of the above procedures, first introduced by ***Breiman et al.*** in 1986 **[14, 15]**. Integrating several data sets from different sources is a major task in DM for more significant discovery, but it may generate ***missing values***. In this case, CART could be a very useful tool, because it involves an elaborate technique to treat the missing values. In fact, CART prepares several surrogate variables for the missing values for each node in the tree so that it might be good for the cases of integrating different set of data records where some uncommon features exist **[16]**.

The splitting criteria (analogous to the normalized information gain for $C_{4.5}$) often used in CART are the ***gini inquiry*** for classification trees and the ***variance reduction*** for regression trees **[12]**.

4. Applications of the C4.5 and CART Algorithms

The C4.5 algorithm is usually recommended for better accuracy and it has been used in a wide range of areas like finance, engineering, etc. **[9, 10, 17]**. CART on the other hand is recommended for the cases of the existence of many missing values and has been favored mostly in medicine domain **[15]**. However, although C4.5 has been neglected in general in medicine applications, one could not claim that CART is always the best algorithm for handling data sets in medicine domain**.** In fact, the resulting DTs are heavily dependent on the available training data sets, and no exceptions from this could happen in medicine domain

For instance, two classifiers based on C4.5 algorithm have been developed in **[17]** and have been applied successfully on medical data for thrombosis diseases.

Moreover, Table 1 **[18]** shows the accuracy of C4.5 and CART for thirteen different medical data sets that have been derived from the UCI machine learning repository [3].

Table 1: The accuracy of C4.5 and CART for each data set

Data set	Accuracy of C4.5	Accuracy of CART
Breast tissue	66.04	**70.75**
Bupa	68.7	**70.14**
Cardiotocography	**98.78**	98.1
Cleveland heart disease	**55.78**	53.14
Hungarian heart disease *	**68.71**	52.04
Switzerland heart disease	**29.27**	21.95
VA Long heart disease	**34.0**	30.0
Dermatology	93.99	**94.81**
Fertility *	**87.0**	55.0
ILPD	**69.64**	64.32
Mammographic mass	82.31	**83.56**
Parkinson's	80.51	**84.62**
Vertebral column-2classes	**81.61**	77.42
Vertebral column-3classes	81.61	**84.52**
No. of wins	8	6

Observing the results of Table 1, one can see that in many cases the performance of CART is worse than that of C4.5, especially for the 'Hungarian heart disease' and for the 'Fertility' that are indicated for emphasis by '*' . On the other hand, in the cases where the CART's accuracy is better, the difference between the two algorithms is relatively small. This suggests that the C4.5 algorithm may generate more reliable results than CART in certain medical applications.

In concluding, more objective results that compare the performance of the two algorithms empirically for a wide variety of medical data sets are needed for a more effective utilization of those two DT algorithms.

5. Conclusion

In this work we have studied the DT algorithms for DM, among which the most popular are the C4.5 and the CART algorithms. The former, being an extension of the Quinlan's earlier ID3 algorithm, is recommended for better accuracy and it has

been applied in a wide range of areas like finance, engineering, etc. The latter, introduced by Breiman et al., is an umbrella used for both classification and regression trees involving an elaborate technique to treat the missing values. Although CART has been favored mostly in medicine domain where the C4.5 algorithm is neglected in general, one could not claim that CART is always the best for handling data sets in medicine applications and further empirical research is needed on this subject for a better utilization of the DT algorithms

References

[1] V. Podgorelec, P. Kokol, B. Stiglic, I. Rozman (2002), Decision trees: an overview and their use in medicine, *Journal of Medical Systems*, Kluwer Academic/Plenum Press, 26 (5), 445-463.

[2] S.B. Kotsiantis (2007), Supervised Machine Learning: A Review of Classification Techniques, *Informatica*, 31, 249-268.

[3] A. Frank, A. Suncion (2010), *UCI Machine Learning Repository*, Irvine, CA: University of California, School of Information and Computer Sciences, available at: http://archive.ics.uci.edu/ml

[4] Y. Gao (2015), Information Mining for Big Information, in W. Pedrycz & S.M. Chen (Eds.), *Information Granularity, Big Dataand Computational Intelligence*, Studies in Big Data, 8, pp. 23-38, Springer.

[5] M. Lenzerini (2002), Data Integration: A Theoretical Perspective, *Symposium on Principles of Database Systems*.

[6] CART Classification and Regression Tree, available at: http://www.salfordsystems.com/products/cart

[7] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z. Zhou, M. Steinbach, D. J. Hand, D. Steinberg (2006), Top 10 Algorithms in Data Mining, *Knowledge and Information Systems*, 14(1), 1-37..

[8] J. R. Quinlan (1986), Induction of Decision Trees, *Machine Learning*, 1, 81-106.

[9] Z. Chang (2011), The application of C4.5 algorithm based on SMOTE in financial distress prediction model, *Proceedings of 2nd International Conference*

on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), pp.5852-5855.

[10] S. Gao (2012), The Analysis and Application of the C4.5 Algorithm in Decision Tree Technology, *Advanced Materials Research*, 457-458, 754-757.

[11] C. E. Shannon (1948), A mathematical theory of communications, *Bell Systems Technical Journal*, 27, 379-423 and 623-656

[12] Wikipedia, *Decision tree learning* retrieved from: https://en.wikipedia.org/wiki/Decision_tree_learning on February, 2018

[13] I.H. Witten, E. Frank, M.A. Hall (2011), *Data mining: Practical machine learning tools and techniques*, 3d Edition, Morgan Kaufmann, San Fransisco, p. 191.

[14] Breiman, Leo; Friedman, J. H., Olshen, R. A., Stone, C. J. (1984), *Classification and regression trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

[15] R.J. Lewis (2000), An Introduction to Classification and Regression Tree (CART) Analysis, *Annual Meeting of the Society for Academic Emergency Medicine*, San Francisco, available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.95.4103&rep=rep1&type=pdf

[16] M.J. Beynon (2009), The Issue of Missing Values in Data Mining, *Encyclopedia of Data Warehousing and Data Mining*, Chapter 171, pp. 1102-1109.

[17] A.- B. M. Salem & A.M. Mahmout (2003), A Hybrid Genetic Algorithm – Decision Tree Classifier, in M.A. Klopotek et al. (Eds.), Intelligence Information Processing and Web Mining, Springer – Verlag, Berlin, Heidelberg

[18] H. Sug (2013), Data Mining in Medicine Domain Using Decision Trees - The Case of CART and C4.5, *Mathematics and Computers in Contemporary Society* (WSEAS Proceedings of the 12th International Conference on Computational Intelligence, Man-Machine Systems and Cybernetics), pp. 212-215, Nonjing, China.

Authors

Dr. Abdel-Badeeh M Salem is currently an Emeritus Professor of Computer Science at Ain Shams University, Cairo, Egypt. His research includes intelligent computing, knowledge-based systems, biomedical informatics, and intelligent e-learning. He has published about 400 papers in refereed journals and conferences. He has been involved in about 500 conferences and workshops as a Keynote Speaker, Scientific Program Committee, Organizer and Session Chair. He is a member of many national and international informatics associations.

Dr. Michael Gr. Voskoglou is currently an Emeritus Professor of Mathematical Sciences at the School of Technological Applications of the Graduate Technological Educational Institute of Western Greece in the city of Patras. He is the author of 14 books and of more than 450 papers published in reputed journals and in proceedings of conferences of about 30 countries in the five continents around the Globe, with very many citations from other researchers. He is the Editor- in- Chief of the "International Journal of Applications of Fuzzy Sets and Artificial Intelligence", a reviewer of the American Mathematical Society and member of the Editorial Board or referee in many international mathematical journals. His research interests include Algebra, Markov Chains, Fuzzy Sets, Artificial Intelligence and Mathematics Education.